DATA PRESENTATION AND INTERPRETATION REQUIREMENTS FOR GEOCHEMICAL EXPLORATION

Stephen Amor
Consultant, Burlington, Ontario
SteveAmor@compuserve.com / www.dirtbagger.com

This section of the workshop will deal with processing single and multiple-element analytical data sets, and integrating the results with field observations and other data, to detect responses related to mineralization, and also as an aid to geological mapping where exposure is poor or where geological maps of desired quality are unavailable for other reasons.

The careful analysis of geochemical data, using standard computer software packages, is an important and affordable way of adding value to an exploration company's assets. The cost of a thorough data review is normally less than ten per cent of total geochemical program costs.

However, with standard computer software, sophisticated methods of data analysis have become very easy to apply, and also abuse. They do not issue warnings when the results they present are based on inadequate or unsuitable data, or when the necessary conditions for their proper application have not been satisfied; nor do they have a requirement that their attractively-packaged conclusions should make geological sense.

The methods of data analysis that will be presented are not intended as a means to avoid careful geological thought, and any conclusions that arise from them must not fly in the face of geological and other observations. On the other hand, they need not necessarily confirm previous inferences (however dearly cherished) from geological and other lines of investigation.

An Underutilized Resource

Reliable multielement analyses are requested increasingly in routine exploration programs, but many geologists who use geochemical methods in exploration are unaware of the potential uses to which these data may be put. Consequently, such data sets tend to be underutilized.

Whether or not they are subjected to sophisticated statistical treatment, multielement analyses constitute a valuable resource of information. The elements offered in common analytical packages may include pathfinders for gold and other mineralization; indicators of felsic, mafic and ultramafic igneous rocks (Figure 1), indicators of calcareous rocks and monitors of environmental conditions (Figure 2).

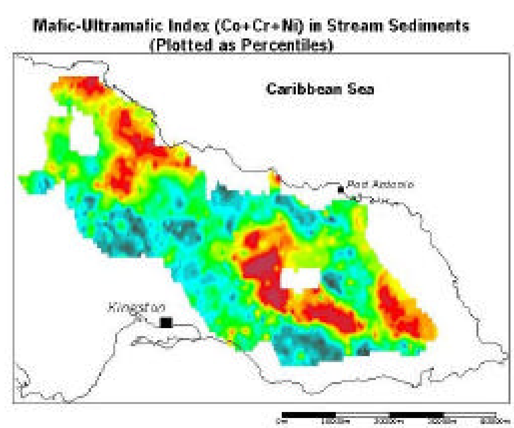


Figure 1: Distribution of "Mafic-Ultramafic Index" in Stream Sediments, Eastern Jamaica

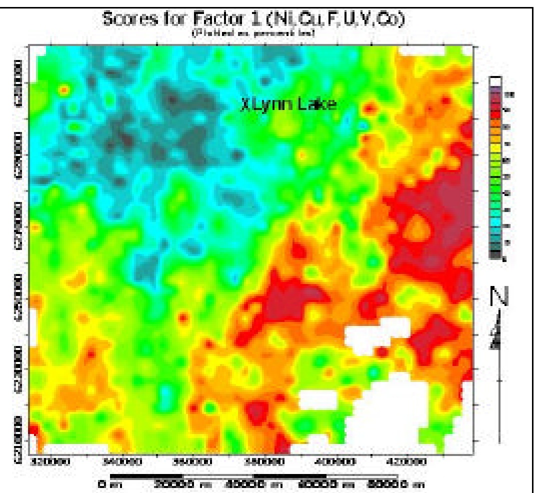


Figure 2: Distribution of interpreted "Lake Agassiz Clay" Factor in lake sediments, Granville Lake Sheet, N. Manitoba

Checking Data Integrity

The truth of the expression "garbage in, garbage out" is nowhere more vividly demonstrated than in data analysis. A significant proportion of the time expended on any interpretation exercise is (or should be) taken up with checking the integrity of the data set. Typically, a file of geochemical data may contain shortcomings such as missing analyses recorded as zeros; "undetectable" values expressed as zeros, inequalities or negative numbers; displaced or mixed columns of analyses and coordinates, and quality-assurance data that has not removed from the main data set for separate evaluation.

Data Transformations

Many statistical techniques are *parametric*; that is, they require the fulfilment of certain assumptions about the data for any conclusions derived from them to be valid. The most important of these is that the frequency distribution of every variable conform to a Normal distribution. Furthermore, the distribution should be free of outliers, and less than 30% of the analyses for each element should be exceeded by the analytical detection limit. This problem, known as *censoring*, is not so serious in the calculation of non-parametric statistics like percentiles, which are a very powerful method of dealing with univariate geochemical data.

It is therefore, necessary to inspect the frequency distributions of the elements in a geochemical data set, and apply appropriate transformations prior to data analysis, or in some cases to omit certain elements altogether. This apparently tedious process can be expedited with good organization, and practice.

Apples and Oranges

Just as it is not meaningful to generalize about the acidity of apples and oranges collectively (since the latter are, in general, considerably more acid than the former) it is not meaningful to generalize about the Cr content of a suite of volcanic rock samples if it is known that the volcanic rocks range in composition from basalt through rhyolite.

The response to mineralization is in many cases very subtle, compared to much stronger responses due to differences in the lithology from which the sample material is derived, or the position in the soil profile from which the sample was collected, to name but two. The latter problem is particularly acute in the thick lateritic profiles in which so much exploration has been carried out over the last few years.

In these circumstances interpretation is aided greatly if the collectors of the samples make and record certain key characteristics of the material they are sampling and the site from which it is collected. This need not be a difficult or time-consuming exercise and literate, locally-recruited employees can be trained to carry it out. When interpretation takes place, the data set can be broken down into major lithological, pedological and other subsets. If information regarding mapped geology is available, the principal source lithology associated with each sample can be incorporated into the database at a later stage, prior to interpretation. The removal of the very strong geochemical signal associated with variable source lithologies and soil types enables more subtle signals, possibly arising from mineralization, to be detected more readily.

The Identification of Anomalies

The principal aim of geochemical exploration remains the detection of variations in the composition of naturally-occurring materials related to the presence or proximity of a potentially economic mineral deposit. Such responses are best recognized during routine exploration if they can be measured beforehand in material, of the type that will be sampled in the main survey, in the vicinity of a known mineralized occurrence of the type sought. This is the essential function of an *Orientation Survey* which has no equal as a means of establishing cutoffs or *thresholds* by which the economic significance of geochemical data can be assessed.

In the common, often unavoidable absence of orientation information it is necessary to examine the data themselves for unusual behaviour and this leads to the concept of the *Anomaly*. Various ways of identifying geochemical anomalies, for which the more cumbersome term "areas meriting follow-up" is nevertheless more precise, will be described and ranked as to their effectiveness. An example of an anomaly map, based on published data and suitable for first-pass identification of follow-up targets, is shown in Figure 3.

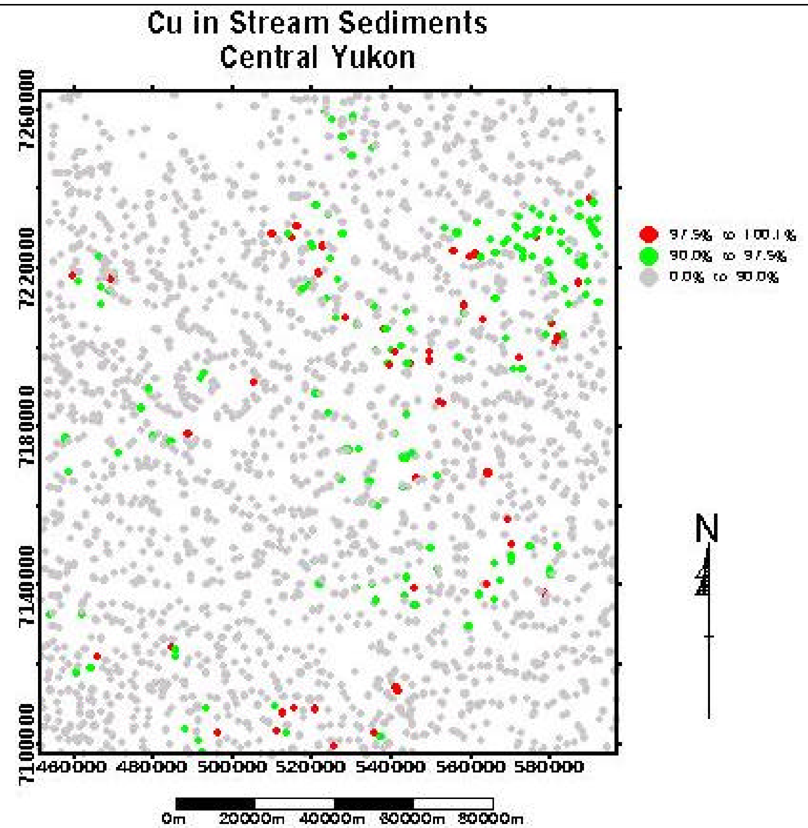


Figure 3: Cu Stream Sediment Anomaly Map, Central Yukon

Red circles indicate values exceeding 97.5-percentile (140 ppm Cu); green circles indicate values exceeding 90-percentile (84 ppm Cu). Data from GSC Open File 2175

Multielement Indices

Methods exist for dealing with multielement data that do not involve multivariate statistics in the strict sense of the term. The calculation of multielement indices is an example of how the element associations described above can be applied to optimize the response to certain mineralization types, or important lithologies. An example of a map showing the distribution of such an index was given in Figure 1.

A spreadsheet application enables these calculations to be readily performed on large data sets. To deal with the “apples-and-oranges” problem described above, the values of each variable can be normalized or converted to percentiles of the overall population, or of subpopulations based on observable criteria such as mapped geology or regolith type.

While indices are useful in many situations, they are an example of an alluring method that can lead to meaningless or even misleading results, if attention is not paid to geological and geochemical realities.

Multivariate Statistical Methods

Computer processing is mandatory in the application of most multivariate statistical methods as it would be prohibitively difficult and time-consuming to undertake it manually. These techniques can be usefully applied without detailed understanding of the underlying mathematics, although some understanding of how they work (and the circumstances under which they do not) is advantageous. The key to the understanding of many of these apparently complex methods lies in the graphical demonstration of a 2D (bivariate) situation and its intuitive extension into higher dimensions (“hyperspace”).

The field of multivariate analysis is very broad, and the discussion of every one of the techniques can be taken to the point where almost every geologist's mathematical ability is inadequate to deal with it. Furthermore, most multivariate software packages come with a variety of options which make it difficult for a new user to know how to proceed. It is important that the user have some knowledge of how a statistical technique works, if only so that situations may be identified where it will not do so.

When deriving any kind of summary statistics, multivariate or otherwise, from a large data set it is important to decide whether the feature sought constitutes a statistical rarity in the sampled population. In a regional drainage survey, for example, the major controls on each sample's composition are likely to consist of gross lithological and surficial or environmental agencies. The predominant element associations (factors), sample associations (clusters) or inter-element and intersample relationships of other kinds, like regression equations, are likely to reflect these controls, and unlikely to reveal much about mineralization, if its presence is manifested in only a few samples. On the other hand, in the follow-up survey of a previously-defined anomaly, the presence or proximity of mineralization is more likely to exert a discernible influence on the data as a whole.

Regression Analysis

In the case of simple linear regression, a set of bivariate data, expressed graphically as an X-Y plot, is fitted with a straight line, that may or may not pass through the origin. This line represents the best estimate of the relationship between what is termed the *dependent* variable (which is normally plotted on the y-axis) and the *independent* variable (x-axis) though no cause-and-effect relationship need be implied. Polynomial regression involves the fitting of a curve, rather than a straight line, to the scatterplot, while multiple regression involves the admission of more than one independent variable; this is analogous to fitting a surface, rather than a line, to a set of points in three (or more) dimensions.

While the observation that a relationship can be established between two geochemical variables may be of academic interest, the principal advantage of regression analysis is in the isolation, in each sample, of the residual component of the dependent variable that cannot be predicted from the independent variable(s). A positive residual value indicates that the dependent variable is higher than predicted, while a negative value indicates that it is lower.

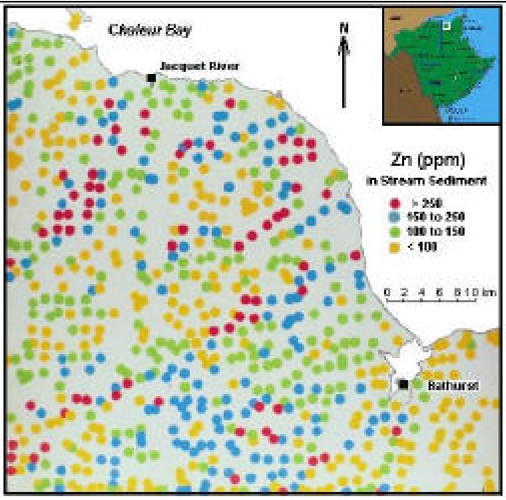
R-mode Factor Analysis

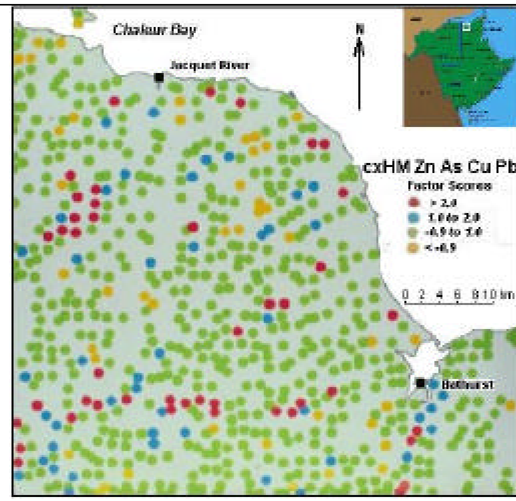
A typical multielement data set may consist of analyses for up to 30 different elements, but it is unlikely that these elements were emplaced by 30 different element-specific processes. Furthermore, the amount of a particular element in a sample is unlikely to be the result of only one process acting on the sample material. The strength of the correlations between certain elements in most naturally-occurring media bears witness to this.

Factor Analysis is a general term given to a variety of related techniques which seek to identify a limited number of controls on a much greater number of observational variables. These are modelled in the form of linear combinations of those variables, termed *factors*. In geochemistry, it is reasonable to suppose that such factors will be more closely related to the processes that have acted on the medium in question, than are the individual elements. Unlike the multielement indices described above, the relative importances or *loadings* of the individual elements in the factors are determined empirically from the data, rather than from preconceived notions regarding their importance. However, the significance of a particular factor may be interpreted *subsequently* in the light of the relationship between its characteristic elements, and the natural processes under which they are known to be mobile.

For each factor, a *factor score*, quantifying the influence of the factor in each sample, can be calculated. Factor scores can be plotted and contoured like any geochemical variable and it is often from their areal distribution that the most useful conclusions can be made.

Figure 4 shows the distribution of Zn in stream sediments from the Bathurst District of New Brunswick, based on data released by the Geological Survey of Canada. When factor analysis was applied to these data, Zn showed strong loadings in three factors and it was concluded that three principal agencies were responsible for contributing the metal to the stream sediment: bedrock lithology (not shown here) whose representative factor was strongly loaded in Ni, Cr, Ba, Co, Zn, Cu, and Mn; mineralization of VHMS affinity (factor loaded in cold-extractable heavy metal, Zn, As, Cu and Pb; Figure 5); and replacement-type mineralization (Ag, Pb, As, Zn, Cu; Figure 6). Distinguishing between the effects of these agencies was not possible on the basis of the Zn results alone.





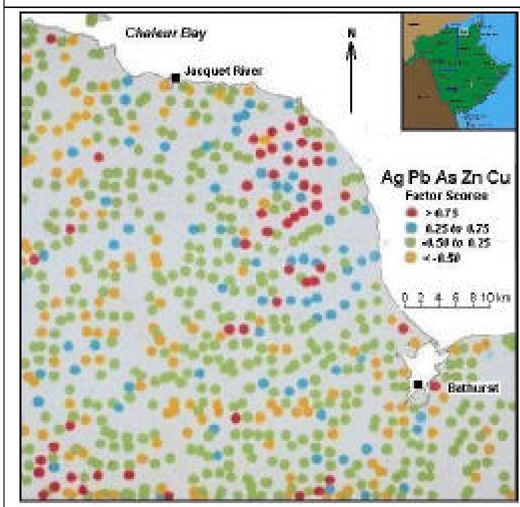


Figure 4 (above left): Distribution of zinc in stream sediments, Bathurst District, New Brunswick

Figure 5 (above right): Distribution of cxHM (cold-extractible heavy metal)-Zn-As-Cu-Pb factor in stream sediments, Bathurst District, New Brunswick

Figure 6 (left): Distribution of Ag-Pb-As-Zn-Cu factor in stream sediments, Bathurst District, New Brunswick

It is stressed that the factors themselves only rarely model the mineralization process, or the dispersion of its products, unless the data are from a detailed follow-up survey. However, a combination of factor and regression analysis can separate the component of an element's concentration level that is attributable to the processes modelled by the factors, from the *residual* component that cannot; this latter component may, in a few samples, be related to mineralization.

Discriminant Analysis

Most geochemical variables are measured on a continuous ratio scale. However, their ultimate function for the explorationist is as an aid to converting them to categorical variables, which may have as few as two values (do we follow this up, or not?) or more (from which lithology does this sample appear to be derived?).

These problems can be dealt with by discriminant analysis. This method is applied to situations where there are previously-defined "training sets" representing classes which differ in some important, observable and important characteristic. From the multivariate observations that make up these training sets, a series of *discriminant functions* are derived, one per defined class. Solution of the functions for the data on a single sample yields an series of indices known as *discriminant scores.* The class whose discriminant score is highest is the one to which that sample is assigned.

The method is useful in two-group situations where suitable training sets are available and it is necessary to discriminate and classify "mineralized" and "unmineralized" or "altered" and "unaltered" samples, where these characteristics are not observable directly in routine samples. The method is also applicable where more than two groups have been identified (for example, when multiple lithologies are present within the unmapped area of a regional stream- or lake-sediment survey).

Conclusion

The primary aim of geochemical exploration remains the direct detection of compositional responses, in naturally-occurring media, to bodies of economic mineralization. However, the aim of this presentation is to demonstrate that careful examination of geochemical data, using techniques of data analysis readily available in commercial software packages, may be used cost-effectively not only to identify these responses directly but also to extract "signals" related to mineralization from the "noise" arising from other sources, and also to detect favourable lithologies, structures and geochemical provinces.

The **Identification of Anomalies** involves filtering out samples whose analytical values of key elements have a reasonable probability of being associated with mineralization. Those samples returning (normally) lower values are not followed up. The cutoff value that determines this separation must have a certain degree of flexibility, and the spatial association of the samples should also be taken into account.

Multielement Indices are linear combinations of analytical variables, weighted at the user's discretion, that are considered to be related to mineralization, or to favourable lithologies, on the basis of previous observation. Because of this built-in property of "making geological sense", and because their derivation is readily explained without resorting to arcane mathematics (although greatly assisted by the availability of computers), indices should always be considered when processing a large multielement data set.

The reduction of "geochemical noise" unrelated to mineralization may be accomplished if certain elements in the data set are strongly influenced by it. **Regression Analysis** enables the modelling and quantification of the noise by such elements and the removal of its effects, thereby emphasizing the more subtle contributions of other agencies which may include mineralization.

Factor Analysis involves the reduction of a data set of many variables into one of considerably fewer linear combinations of those variables, that account for an acceptable proportion of the total data variance and which can often be more readily related to recognisable geological and environmental processes than the input variables themselves.

Finally, **Discriminant Analysis** enables the assignment to previously-defined classes (e.g. mineralised vs. unmineralized) of samples of unknown provenance, based on their multielement composition.

All valid statistics are the result of the condensation of larger amounts of raw data into relevant, concise and usable information. The successful application of data analysis depends on the efficiency of this condensation process. The decisions based on the information may involve millions of dollars and it is desirable that they are not "diluted" by irrelevant considerations (undercondensation) or that relevant relevant information is "boiled away" (overcondensation). The risk of either of these outcomes may be reduced by a visual, non-statistical **Check of Data Integrity**; selection of appropriate **Data Transformations** prior to application of parametric methods; and in compensation of previously-identified inhomogeneities in the data (the **Apples and Oranges** problem).